УДК 656.7.085:657.71(045)

ІНЖЕНЕРНО-ЛІНГВІСТИЧНІ ПРИНЦИПИ АНАЛІЗУ ТЕКСТІВ

О. Є. Литвиненко, Д. А. Бурко
Національний авіаційний університет
litvinen@nau.edu.ua

Виділено основні процеси лінгвістичного аналізу текстів. Доведено, що основними одиницями мови є словосполучення, а зміст словосполучень, як правило, не зводиться до змісту слів, що входять до їх складу. Наведено ефективні методи процесів семантичної, морфологічної, синтаксичної обробки текстів.

The article deals with determination of the main processes of linguistic text analysis. It is asserted the basic units of language are words and phrases, but their content is not reduced to the words' one that belongs to their structure. There are proposed the effective methods of semantic, morphological and syntactical word processing.

Вступ

У сучасному інформаційному світі одним із актуальних напрямів розробки штучного інтелекту стала комп'ютерна лінгвістика. На сьогодні розроблено велику кількість електронних словників, лексичних систем, програм машинного перекладу, систем анотування та реферування, створено програми, які самостійно пишуть твори не лише у прозовій формі, а й у віршованій.

Людство одержало можливість користуватися лінгвістичними розробками чи-то для перекладу, чи-то для пошуку інформації в Інтернеті, чи-то в інших галузях. Проте робота таких систем із текстами приховує в собі низку невирішених питань:

- якими насправді є закони створення зв'язків у текстах?
- чому випадковий набір слів наш розум розпізнає як осмислений?

Дослідження

Дослідження в цій галузі тісно пов'язані із загальними законами асоціативного мислення людини. А мислення, як відомо, відображає закони світу, в якому спорідненість речей віддзеркалюється у наборі слів.

Коли ми скажемо: «Потяг, квиток, валіза…», то розум буде намагатися відтворити ситуацію, в якій дані об'єкти будуть заходитися у певній взаємодії між собою. Це може бути ситуація на зразок «Початок літнього відпочинку, провідник потягу перевіряє квиток пасажира, і допомагає занести валізу до вагону».

Наш мозок уміє співставляти прості слова і образи один з одним, таким чином зв'язний текст впливає на мислення людини і викликає послідовність уявлень знайомих ситуацій. Але така послідовність неоднозначна і залежить від досвіду людини і ступеня її інтелектуального розвитку.

Узагальнюючи сказане, варто звернути увагу на деякі закономірності, які використовуються при розробці алгоритмів аналізу текстів. Одну з них свого часу було виділено німецьким ученим Джорджем Зіпфом. Він помітив, що «Слова с большим количеством букв встречаются в тексте реже коротких слов» [7], на основі чого вивів два закони.

Перший закон зв'язує частоту появи слова в тексті з рангом цієї частоти. Словам, які зустрічаються найчастіше, присвоюється ранг 1, а тим, які рідше, — 2 і т. д. Д. Зіпф виявив: якщо помножити ймовірність наявності слова в тексті на ранг його частоти, то отримаємо приблизно сталу величину. Така залежність може бути представлена гіперболою. Ці дані свідчать про те, що, якщо найпоширеніше слово буде зустрічатися в тексті 100 разів, то наступне за поширеністю — приблизно 50 разів. Так, найпоширенішими словами в українській мові можна назвати: і, а, що, на, у, в, з, бути, по, я, він, який (рис. 1).

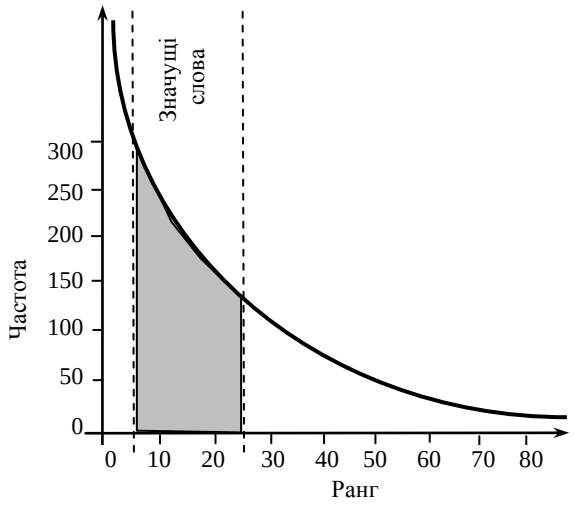

Рис. 1. Перший закон Зіпфа

Другий закон Д. Зіпфа констатує, що частота і кількість слів, які входять до тексту з цією частотою, пов'язані між собою. І якщо побудувати графік, відклавши на осі *X* частоту введення слова, а на осі *У* — кількість слів, які входять у текст із певною частотою, то отримаємо криву, яка збереже свої параметри для всіх текстів, написаних людиною (рис. 2).

Сутність законів Д. Зіпфа дає підстави стверджувати, що кожна мова має слова, які зустрічаються частіше, ніж інші, але не мають значення. І є слова, які вживаються рідше, але мають велике смислове навантаження. Це й представлено на графіку (див. рис. 1). Слова, які зустрічаються занадто часто, — це займенники й прийменники,

а які дуже рідко — в більшості випадків не несуть особливого смислового навантаження.

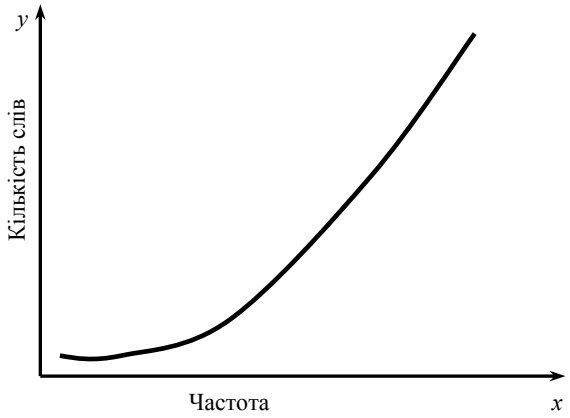

Рис. 2. Другий закон Зіпфа

У сучасному світі, кожна пошукова система у своїй основі використовує закони Д. Зіпфа. Вона вирішує, які слова вибрати як найбільш важливі в тексті. Для того, щоб система зробила такий вибір, необхідно «навчити» її аналізувати тексти. І саме на цьому шляху ми зустрічаємося з деякими труднощами.

Як стверджує вчений А. В. Анісімов, найбільша складність у їх розв'язанні полягає в тому, що існує «інформаційна ізольованість процесів обробки на кожному етапі аналізу — під час роботи процесу обмін даними з іншими процесами не відбувається» [1]. Семантичний, синтаксичний та морфологічний аналізи тексту, що здійснюються людиною, є паралельними взаємодіючими процесами. При визначенні структури речення один процес використовує результати інших. Але напрямок такої взаємодії є не стільки «знизу — вгору» (морфологія визначає синтаксис, синтаксис — семантику), скільки «згори — донизу» — семантика керує синтаксисом та морфологією, синтаксис має вплив на морфологію. Графічно це показано на рис. 3.

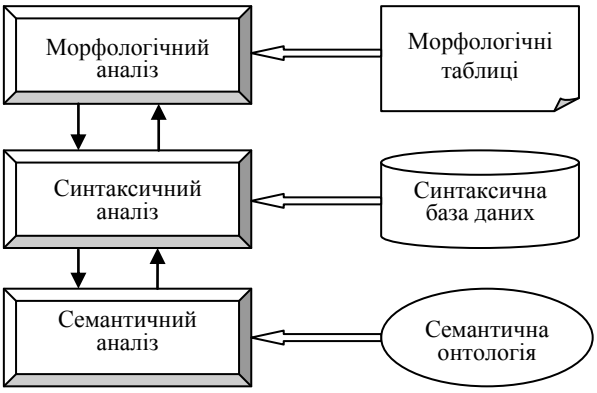

Рис. 3. Блок-схема аналізу тексту

Коли в результаті аналізу один процес зустрічає неоднозначність, він підключає процес вищого рівня, який намагається ефективно розв'язати цю неоднозначність. Звідси випливає, що моделі процесів аналізу текстів доцільно розробляти як систему паралельно розподілених процесів із заданим відношенням підпорядкування.

На нашу думку, необхідно розглянути процес морфологічного аналізу тексту: словоформу розбити на морфеми (за умови, якщо таке розбиття можливе) шляхом пошуку морфем у спеціальних списках, де кожній морфемі відповідатиме певна інформація; потім із інформацій до морфем побудувати загальну інформацію до словоформи.

У процесі морфологічного аналізу виділимо такі основні моменти, що є спільними для різних мов та алгоритмів:

1. Пошук словоформи (або залишку) в певних списках морфем (у таблицях). Під пошуком мається на увазі послідовне порівняння текстової одиниці (словоформи або залишку) з елементами списку (з основами, префіксами або суфіксами) доти, доки елемент списку політерно не «вкладеться» зліва направо до текстової одиниці, що обробляється.

2. Вибір потрібного елемента з кількох можливих. Тут мається на увазі випадок, коли до текстової одиниці одночасно можуть «вкладатися» різні морфеми (омонімічні або спряжені). Необхідний вибір робиться на основі спеціальних позначок, що приписані морфемам та відображають їхню сполучуваність.

Так, за допомогою поміток при основі обирається правильний суфікс (із декількох, що «вкладаються» до цього закінчення); при розтинанні залишку на суфікси може бути виправлено помилку, яку допустили раніше (під час відтинання основи) і т. д.

3. Вибір інформації про окремі морфем та об'єднання цих інформацій у загальну інформацію до словоформи, що аналізується.

4. Виявлення індивідуальних особливостей морфем (якщо такі особливості є) та врахування впливу цих особливостей на загальну інформацію про словоформи.

5. Вибір подальших дій після того, як завершено обробку наступної морфеми. Послідовність дій зазвичай визначається властивостями щойно обробленої морфеми.

Конкретний зміст цих п'яти основних моментів морфологічного аналізу змінюється залежно від алгоритму, але самі ці моменти залишаються майже незмінними.

Основні моменти морфологічного аналізу, що наведені вище, дають змогу чітко його осмислити.

У загальному вигляді алгоритм морфологічного аналізу складається з п'яти частин.

1. Загальні правила є основною частиною алгоритму (власне алгоритмом) і набором правил, що визначають послідовність операцій та взаємодію решти частин.

2. Список суфіксів містить перелік суфіксів мови разом із деякими вказівками, що необхідні для їх правильної обробки.

3. Список інформацій про суфікси містить перелік основної інформації про ті суфікси, що містяться у другій частині (інформація, яку слід перенести до інформації про словоформи).

4. Нестандартний запис є набором указівок про індивідуальні особливості суфіксів співвідносно до особливостей основ, а також про те, як ці особливості мають бути враховані.

5. Таблиця, в якій наведені випадки, коли повністю збігаються деякі форми від різних основ (при тому, що інші форми від цих слів — різні).

Загальну схему морфологічного аналізу зображено на рис. 4.

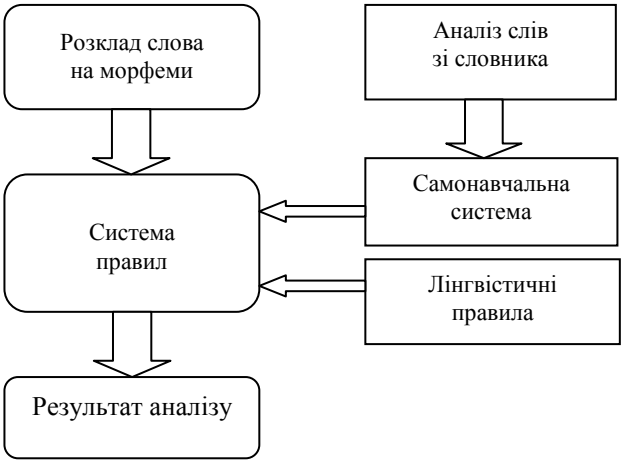

Рис. 4. Нечіткий морфологічний аналіз

Особливо важливо вказати на синтаксичний аналіз. Передусім необхідно виділити основні проблеми синтаксису: проблема його об'єкту, співвідношення із семантикою й морфологією; проблему слова, групи, фрази як синтаксичних одиниць, а також проблему основних понять синтаксису: відношення (зв'язку), функції, структури, формальних показників.

Завдання синтаксичного аналізу полягає в тому, щоб, використовуючи морфологічну інформацію про словоформи, одержану на морфологічному етапі, побудувати синтаксичну структуру вхідного речення. Об'єктом аналізу є речення, яке до моменту синтаксичного аналізу подається у вигляді інформаційних ланцюжків до словоформ. Результатом аналізу є синтаксична структура речення, представлена як сукупність даних про синтаксичні зв'язки між його одиницями.

Перейдемо до методів синтаксичного аналізу текстів. Учений К. Селезньов розділив їх на дві підгрупи: методи з фіксованим набором визначених наперед правил та самонавчальні методи. Правила представлені не в класичному вигляді (якщо… , то…), а у вигляді граматик, які задають синтаксис мови.

Можна погодитися з твердженням К. Селезньова, що аналіз на основі самонавчальних систем є поки що мало вивченою сферою. Він полягає в такому: розробляється низка прикладів, які містять у собі початкове речення і результат його аналізу. Цей результат вводиться людиною, яка займається «навчанням» системи, у відповідь на кожне вхідне речення. Якщо до системи буде введено речення, що не має відповідних прикладів, то вона сама згенерує результат. Для реалізації цієї системи використовуються такі методи: нейромережі, дерева виходу та методи пошуку найближчого сусіда.

У результаті можна дійти до висновку про те, що вирішення завдань синтаксичного аналізу може стати основою для більш удосконалених синтаксичних коректувальників та для побудови алгоритмів якісного семантичного аналізу.

Наступним процесом текстового аналізу є семантичний аналіз, завдання якого полягає у визначенні семантичної структури речень та тексту в цілому. Умовно його можна поділити на три етапи.

Завдання першого етапу полягає в знаходженні в базі семантичної онтології концепту, який відповідає конкретному значенню слова чи словосполучення. Другого — у побудові семантичного фрейму речення.

Третя фаза смислового аналізу включає в себе об'єднання ізольованих семантичних фреймів речень у зв'язну семантичну мережу тексту. Об'єднання двох структур в одну мережу відбувається згідно з принципом об'єднання семантично тотожних вершин, тобто якщо вершини посилаються на один концепт, вони об'єднуються в одну вершину.

У процесі проходження цих етапів система створює гіперграф, кожен вузол якого має ім'я — слово, яка характеризує сенс вузла. Семантична відстань між двома вершинами (концептами) може бути проінтерпретована як довжина найкоротшого шляху між відповідними вершинами у графі онтологічної мережі. Можна запропонувати два шляхи знаходження цієї відстані:

1. Простий пошук шляху: типи зв'язків—відношень не враховуються; вважається, що всі дуги одного типу. Ще одним варіантом цього підходу є числове ранжування зв'язків—відношень, де дугам різного типу присвоюються різні вагові коефіцієнти, але сам алгоритм залишається без змін.

2. Евристичний пошук: при побудові найкоротшого шляху дозволяються лише деякі послідовності типів зв'язків-відношень.

Коли найкоротший шлях знайдено, його довжина буде обрана як семантична відстань між певною парою концептів.

Більшість методів семантичного аналізу працюють із сенсом слів. Отже, повинна існувати база, спільна для всіх методів, яка давала змогу виявляти семантичне відношення між словами. Такою основою є тезаурус мови. На математич-

ному рівні він є орієнтовним графом, вузлами якого є слова в їх основній словоформі. Дуги задають відношення між словами і можуть мати ряд відтінків.

Синонімія — слова-синоніми.

Антонімія — слова-антоніми.

Гіпонімія — одне слово є одиничним випадком іншого (наприклад, меблі і стіл).

Гіперонімія — протилежне до гіпонімії.

Еквонімія — слова, які є гіпонімами одного й того ж слова.

Амонімія — слова, що однаково пишуться і вимовляються, але мають різний зміст.

Паронімія — слова, дуже подібні за звучанням, нерідко — й за значенням, але не тотожні.

Конверсиви — слова, що мають протилежний зміст (купив — продав).

Висновки

Розглянуто алгоритми аналізу текстів, а саме: морфологічний, синтаксичний та семантичний, які можуть бути використані при створенні лінгвістичних систем для реферування та індексації текстів, під час розробки діалогових систем між людиною та комп'ютером.

Останнім часом завдяки розвитку системи документообігу, наявності довідників, що постійно оновлюються, та ряду інших факторів, спостерігаємо нагромадження масивів спеціалізованих текстових документів. За аналогією зі структурованою інформацією, коли вдосконалення засобів аналізу призвело до появи баз даних, розвиток документообігу з часом потребуватиме створення повнотекстових баз, які давали б можливість усебічного аналізу текстів.

ЛІТЕРАТУРА

1. *Анісімов А.В.* Алгоритмічна модель асоціативно-семантичного контекстного аналізу текстів природною мовою. // Проблеми програмування / А. В. Анісімов, О. О. Марченко, А. О. Никоненко. — К. — 2008. — № 2—3. — С. 379—384.

2. *Анисимов А. В.* Система обработки текстов на естественном языке // Искусственный интеллект / А. В. Анисимов, А. А. Марченко. — 2002. — № 4. — С. 157—163.

3. *Анисимов А. В.* Компьютерная лингвистика для всех: Мифы. Алгоритмы. Язык / А. В. Анісімов. — К. : Наук. думка, 1991. — 208 с.

4. *Ахо А.* Теория синтаксического анализа, компиляции и перевода: В 2 т. / А. Ахо, Дж. Ульман. — М. : Мир, 1978. — Т. 1. — 612 с.; Т. 2. — 487 с.

5. *Белоногов Г. Г., Быстров И. И., Козачук М. В. Новоселов А. П., Хорошилов А. А.* Автоматический концептуальный анализ текстов // Сб. «Научно-техническая информация», сер. 2. — ВИНИТИ. — М. —2002. — № 10. — С. 26 — 32.

6. *Марченко О. О.* Моделювання семантичного контексту при аналізі текстів на природній мові // Вісник Київського університету. Сер. фіз.-мат. науки. — К. — 2006. — № 3. — С. 230—235.

7. *Шемакин Ю.И.* Компьютерная семантика / Ю. И. Шемакин, А. А. Романов. — М. : Научно-образовательный центр «Школа Китайгородской», 1995. — 344 с.

8. *Джордж Кингсли Зипф* (Електронний ресурс) — http://www.antula.ru/law-zipf.htm.

Стаття надійшла до редакції 12.10.09.